DOCUMENT RESUME

ED 121 827 95 TM 005 258

AUTHOR Coulson, John E.
TITLE Problems and Approaches in Education Program Evaluation.
INSTITUTION System Development Corp., Santa Monica, Calif.
SPONS AGENCY Office of Education (DHEW), Washington, D.C.
PUB DATE [Apr 76]
CONTRACT 300-75-0332; OEC-0-73-0831; OEC-0-73-6336
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1975)

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Compensatory Education Programs; Criterion Referenced Tests; Educational Objectives; Educational Research; Elementary Secondary Education; *Federal Programs; *National Programs; Program Effectiveness; *Program Evaluation; Research Problems; Testing Problems; Time Perspective
IDENTIFIERS Emergency School Aid Act

ABSTRACT

Although dealing broadly with issues involving program evaluation, this paper is especially concerned with large-scale evaluations of federally funded education programs that are national in scope and, to some extent at least, compensatory in nature. The paper concentrates on a few basic problem areas that seem inherent in the concept of evaluation. They include, for example, problems in defining treatment and outcome variables that will yield useful evaluation results, and difficulties in establishing comparative benchmarks against which program outcomes should be gauged. The paper concludes with a discussion of the role of evaluation vis-a-vis education research, and suggests a strategy by which future evaluations might make a greater contribution to an understanding of important underlying relationships in the teaching-learning process. To illustrate both problems and approaches in evaluation, examples are given from two large studies currently being conducted by System Development Corporation for the U.S. Office of Education: an evaluation of the Emergency School Aid Act (ESAA) and a multiyear evaluation of the sustaining effects of compensatory education. (RC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED121827

PROBLEMS AND APPROACHES IN EDUCATION PROGRAM EVALUATION

John E. Coulson

System Development Corporation

The work upon which this publication is partially based was performed pursuant to Contracts OEC-0-73-0831, OEC-0-73-6336, and 300-75-0332, with the U.S. Office of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the positions or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred.

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL BY MICROFICHE ONLY HAS BEEN GRANTED BY

System Development Corp.

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Prepared for presentation at the annual meeting of the American Educational Research Association, San Francisco, April 20, 1976.

TM005 258

PROBLEMS AND APPROACHES IN EDUCATION PROGRAM EVALUATION

John E. Coulson

System Development Corporation

Although this paper deals fairly broadly with issues involving program evaluation, it is especially concerned with large-scale evaluations of federally funded education programs that are national in scope and, to some extent at least, compensatory in nature. Rather than cataloging an almost infinite list of potential pitfalls in evaluation, this paper concentrates on a few basic problem areas that seem inherent in the concept of evaluation. They include, for example, problems in defining treatment and outcome variables that will yield useful evaluation results, and difficulties in establishing comparative benchmarks against which program outcomes should be gauged. The paper concludes with a discussion of the role of evaluation vis-a-vis education research, and suggests a strategy by which future evaluations might make a greater contribution to an understanding of important underlying relationships in the teaching-learning process.

To illustrate both problems and approaches in evaluation, examples are given here from two large studies currently being conducted by System Development Corporation for the U.S. Office of Education. These studies are held out not as models of perfection, but rather as representative of some of the larger and more recent evaluations. At the same time, they also reflect the improved quality that seems evident in evaluation contracts conceived and awarded by federal agencies over the past five years.

One of the studies used illustratively in this paper is an evaluation of the Emergency School Aid Act (ESAA). Under the ESAA Basic Grants Program, federal funds are provided to eligible school districts to help them meet special needs incident to desegregation. Schools receiving Basic Program funds have multi-racial enrollments. Activities conducted with these funds include both desegregation-related projects, such as training of staff

Prepared for presentation at the annual meeting of the American Educational Research Association, San Francisco, April 20, 1976.

members in race relations, and efforts to enhance students' academic skills. The ESAA Pilot Program, by contrast, is directed toward schools with large numbers of minority students (at least 50%). In Pilot award schools the ESAA funds are used entirely for compensatory education purposes, particularly the improvement of students' reading and mathematics skills.

The ESAA Evaluation Project, which the present author directs, attempts to assess the impact of ESAA funding on student achievement, on students' perceptions of discrimination in the schools, and on the degree of minority group isolation in the funded districts. In addition, the study seeks to identify critical factors in the local ESAA projects that contribute to or interfere with program success. The ESAA evaluation contract is now in its third and final year of data collection. A Year One Report has been issued, and the Year Two Report is currently being reviewed by USOE.

The second study referred to in this paper is a multi-year evaluation of the sustaining effects of compensatory education (including, but not limited to, the Title I program) on basic skills. This study is being conducted by SDC under the direction of Dr. Launor Carter; like the ESAA study, it is funded by USOE. The Sustaining Effects Evaluation is currently in its planning and design phase, with the first of three years of data collection scheduled to begin this fall.

Defining Independent Variables

Most large-scale evaluations deal with programs mandated by Congress or other legislative bodies; increasingly, in fact, the evaluations themselves are built into the programs' enabling legislation. Congress often tends to view programs, even complex national programs such as Title I or Head Start, as unitary entities. Such a viewpoint is entirely reasonable, since Congress oversees many large programs, and must therefore attend more to the relative merits of competing programs than to the internal details of any one program. Basically, Congress wants evidence on which it can base or justify decisions to continue, expand, or phase out entire programs. It asks, therefore, whether a program is having some impact that justifies its continued existence.

Sensible though it may be from Congress' perspective, this rather simple view of program evaluation creates serious problems for the researcher or evaluation specialist, as well as for program practitioners and administrators at the operational level. One problem is that such summative evaluations of total programs are of little help in guiding program revisions. Depending on the timing, in fact, they may nip worthwhile programs in the bud before they have had a chance to demonstrate their true value. Any study that treats a complex program as a homogeneous entity ignores the important variations among different local applications of that program--variations that may, in some instances, be as large as the differences between major programs. Global evaluations may thus obscure the fact that some local efforts are highly successful, while others are actually detrimental.

From the perspective of the researcher or analytically oriented evaluator, perhaps the greatest problem with the overall summative evaluation paradigm is that the program funding/no-funding distinction is extremely distal in its relationship to outcome measures such as student achievement. Students are obviously not directly affected by program funding decisions. To the extent that such decisions have an impact, it must be through program variables having more proximal relationships to achievement, such as teacher behaviors, availability and use of textbooks and other instructional media, student interactions with advisors, etc. These proximal variables are, in turn, influenced by the total district and school resources, not just by the program being evaluated. Thus, for any serious understanding of program effects, there seems a crucial need to delve beneath the surface of the program funding issue, and to examine intervening variables at the local project level.

Fortunately, most large-scale evaluations funded recently by HEW, at least in education, have called for the examination of interactions among program variables, contextual variables, and outcome measures, in addition to an assessment of overall program impact. Both the ESAA and Sustaining Effects studies, for example, involve the collection of data on large numbers of program variables, including teacher-student interaction modes, and the

analysis of relationships between those variables and student outcomes. In a sense, these studies serve formative as well as summative evaluation purposes, and should help to guide future program design efforts.

Several difficult problems remain, however, even after the decision is made to examine program variables at the proximal classroom or school level. When evaluating a program such as Head Start or Title I, it would be helpful to identify certain common patterns or clusters of features that characterize that program and distinguish it somewhat from other parts of the total school complex. Only in this way can the program be regarded as more than a miscellaneous and ideosyncratic collection of local projects. Yet the guidelines for funding of most large-scale educational programs are extremely broad and permit widely differing interpretations as to the specific goals and implementation techniques of local projects. In ESAA we have found extreme variations not only within a district or school, but also among classrooms within a school. Even the in-depth analyses that we conducted in a subset of 30 schools did not immediately reveal any consistent program patterns across schools, and few such patterns within a school. We are still examining the in-depth data, however, and are applying factorial and correlational methods in a post hoc effort to find clusters or profiles of consistently associated program features.

In a few programs such as Follow Through, schools have been asked to adopt one of several well-defined program approaches, or to specify a comparable approach of their own. Such a "planned variations" procedure has distinct advantages from a research viewpoint, since it helps to define a limited number of clearly specified models for comparison purposes. However, placing such constraints on local education agencies might be politically difficult for much larger programs such as Title I.

Identifying Relevant Outcome Dimensions

One of the most obvious requirements for any evaluation, yet paradoxically one of the most difficult things to achieve, is consensus as to the most appropriate measures of program impact. Disagreements about the selection of these measures occur at all levels, from technical arguments about the psychometric properties of specific instruments, to more philosophical

debates about culture-free tests and value-free evaluations. At the heart of most disagreements is the lack of clear-cut, unambiguous goals for most large social programs. Congressional legislation very rarely defines program objectives at a concrete, operational level. Instead, Congress typically uses phrases such as "overcoming cultural disadvantagement," or "helping districts to improve the quality of instruction." It is then left to the Administration to define these goals somewhat more operationally, by means of program guidelines and regulations. Often the goals represented in these guidelines are further explicated and interpreted in a Request for Proposal issued by HEW for an evaluation of the program. Finally, the evaluation contractor selects or develops one or more instruments designed to reflect the program's success in meeting its specified objectives.

Opportunities for disagreement and second-guessing abound at every stage in the translation from broad Congressional mandate to specific instruments. HEW may be accused of misreading Congressional intent regarding the program's true goals. USOE may be attacked as interpreting too narrowly or selectively the outcome dimensions implicit in the program guidelines. The contractor may also be criticized, not only for perceived flaws in USOE's selection of outcome dimensions, but also for its own real or imagined mistakes in designing instruments to reflect those outcomes.

The point here is not that criticisms are necessarily inaccurate or unfair, but that they often represent widely divergent thinking about the goals of the program and of the evaluation. It is sometimes claimed, for example, that the real benefit of Program X is that it pumps money into the schools and aids the local economy. Or, the program may be said to reduce community unrest by demonstrating good faith and genuine concern on the part of the government. But should the contractor attempt to evaluate educational programs on goals of this type, when such goals are almost never explicitly stated by either Congress or HEW? Clearly the time to resolve basic differences of opinion concerning program objectives is <u>before</u> any Request for Proposal is released for an evaluation contract.

Even when there is reasonable agreement on the general program goals and relevant outcome dimensions, there may be serious disputes about the effectiveness of various types of instruments for measuring those outcomes.

Recently, for example, standardized academic achievement tests have been under fire for a variety of reasons, ranging from accusations that such tests are biased against certain subcultures, to objections that the tests do not adequately accommodate wide differences in the priorities of different local projects. Some critics argue strongly that no single test can legitimately be used in a global evaluation of any large, heterogeneous program; the only fair and meaningful procedure in such cases, it is claimed, is to use criterion-referenced tests tailored to the specific goals of each local project in the program. There are two substantial difficulties with this latter approach. First, such a study would be difficult to implement, because of the time and effort that would be required to develop individually tailored tests having the desired psychometric properties, for any large sample of local projects across the country. The magnitude of such an effort can only be appreciated by someone who has attempted to elicit behaviorally-defined program goals from local project personnel. But even assuming that such local goals could be defined and appropriate tests developed, it might be extremely difficult to communicate the results of such an evaluation in terms that would be meaningful to Congress or to other high-level decision makers. For example, what would Congress make of a statement that in 75 percent of the sample projects at least 80 percent of the students had met some locally specified criterion of performance? As stated earlier in this paper, Congress tends to think of Head Start or of Title I as a homogeneous entity; they want to know whether <u>it</u> is achieving its goals. The evidence that Congress is likely to understand and appreciate will probably be normative--for example, do students in the program learn more than students did before the program was introduced, or better than they might have been expected to perform based on national norms. Because the primary mission of our ESAA study and of the Sustaining Effects evaluation is to answer Congress' questions about overall program effectiveness, both studies use standardized achievement tests. However, in the Sustaining Effects study we are also examining the possibility of analyzing each sample project's success with respect to a subset of items in the standardized test; that subset would consist of the items selected by the local program staff as most representative of their project's top priority goals.

Timing the Evaluation

The timing of an evaluation is usually not set by the evaluation contractor, but by the funding agency--for example, USOE. USOE's timing, in turn, is often affected by Congressional reporting requirements, by availability of funding, and by delays inherent in the procurement processes of all governmental agencies. Thus, of the assorted factors influencing the timing of evaluations, some of the most critical may be purely logistical and may have little to do with substantive issues that should determine the timing. Yet those substantive issues are extremely important, and ideally should stimulate careful trade-off analyses to maximize the benefits of each evaluation. One factor that should be considered in such trade-off analyses is the need for timely information. Perhaps the most often cited reason for the failure of evaluations to influence policy decisions is that the results of the evaluations come too late to be of any practical help. Program funding decisions, according to this argument, cannot be suspended for years and years until white-coated researchers produce technically elegant but stale reports. On the other hand, an equally common complaint about program evaluations is that they are premature--that they make judgments about programs before giving those programs a fair opportunity to prove their worth. Rome was not built in a day, and neither are successful programs, especially when those programs are complex, when they must be integrated with ongoing school programs, or when they are attempting to bring improvements in the performance of severely disadvantaged students.

Clearly, both of these arguments have some validity, yet they also have conflicting implications for the timing of evaluations. Any evaluation schedule should certainly be tailored to the specific circumstances of the program in question. In developing the schedule, one should ask the question: "What is the minimum time that any program, however, effective, might require to produce results of the sort we are looking for?" For many educational outcomes, such as significant and generalized improvements in children's basic learning skills, the answer to this question is likely to be at least two years, and probably three or four. Yet, most evaluations require critical analyses of student gains after only one year of program exposure, and in some cases, the year in question includes an initial period

when the program itself is still being organized and coordinated. Some efforts have already been made to modify Congressional and Administration expectations, so that there is a more realistic concept of the time required to show important educational improvements. These efforts have been only partially successful, and much still remains to be accomplished in this regard.

Even when there is agreement to defer the preparation and release of evaluation reports for two or more years, it is highly desirable to begin the collection of outcome data at the earliest possible date. Where possible, in fact, the collection of such data should start before program implementation, but in the same educational setting (e.g., school or district) where the program is to be installed. The use of pre-program and early-program data can be extremely valuable in establishing baseline performance measures, and in tracing performance over successive years of program operation and exposure. For example, a graph showing fall and spring test results over a period of several years, especially if the interval includes a pre-program year, can be much more informative than a statistical analysis of any single pre-post test comparison.

In cases where student achievement or similar outcome measures do not provide a fair basis for evaluating program success in the initial year or two, other types of measures may be used effectively for interim evaluation purposes. Almost immediately, for example, process measures can be collected that describe the program's operations and organization; from these measures, inferences can be drawn concerning the extent to which the program is functioning as originally designed. In the ESAA evaluation, one of the major first-year findings was related to the allocation of ESAA resources at the district, school, and student levels. We found that districts and schools receiving ESAA funds were among the most needy in terms of students' entry skills, and that individual students selected for program participation were consistently among the most disadvantaged in those districts. We also found a significant relationship between the specific areas of need or disadvantagement, such as reading or mathematics, and the proportions of ESAA funds allocated to those areas of instruction. These findings are quite encouraging, and contrast sharply with results of previous studies of certain other large-scale programs targeted toward specific subgroups.

Establishing Criteria for Assessing Program Success

Assuming that the treatment or program being evaluated has been operationally defined, and that the relevant outcome dimensions have been identified, the next major hurdle is to determine what type of results will be accepted as evidence of program success. The question here is not one of technical details, such as whether to use a .01 or .05 level of significance test, but concerns a much more basic issue, namely: What is the benchmark against which the program's performance should be evaluated? In a very real sense, every program evaluation can be thought of as a comparison of that program's outcome with the outcome that might have expected without the program. But, of course, the without-program results can only be indirectly inferred. Every evaluation design, whether experimental or quasi-experimental, represents an effort to get around the fact that you cannot perfectly replicate conditions with and without the program in question. That is, you cannot run the clock back and then repeat the critical interval holding all other conditions identical but this time omitting the program. Though some designs are more successful than others in circumventing this fundamental constraint, all are compromises or approximations to the unobtainable ideal comparison. From this viewpoint, in fact, Campbell and Stanley's (1963) "rival explanations," which they showed to be associated with and threatening the validity of different research designs, reflect the various ways in which each design departs from the ideal comparison.

There are, of course, several variations on the general concept that the ideal goal is to compare observed program outcomes with outcomes that would have resulted without the program. For example, in some evaluations, several specific treatments may be compared with each other and with a null condition involving no treatment exposure. Often the without-program comparison condition is not clearly defined, but is a diffuse aggregation such as a random sample of students from across the country. An example of this latter type of comparison is the use of national test norms. Still another variation involves the comparison of outcomes for several different levels of program exposure or intensity, rather than for a simple program/no-program dichotomy.

It is generally, though not universally, accepted that the closest approximation to the ideal conditions for program evaluation involves the use of the true experimental design, with random assignment of students or schools to treatment and control conditions. But for practical reasons, such a design is rare in large-scale program evaluations. Typically, it is not feasible to determine on a random basis which students are to participate in some innovative or compensatory education program.

Our ongoing evaluation of the Emergency School Aid Act (ESAA) does follow the true experimental paradigm; we have selected pairs of similar schools in ESAA-funded districts, and have randomly designated one member of each pair to participate in ESAA-funded activities, while the other school serves as a non-participating control school. The success of the random assignment procedure is evidenced by the fact that the participating and non-participating schools, and the sample students in those schools, were essentially identical on a wide range of variables at the start of the school year. However, the true experimental design, for all its advantages over most other designs, has not been a panacea in the ESAA study. In fact, that study illustrates rather clearly the gap that can still exist between the true experiment and the ideal program evaluation. In the ideal situation, program outcomes are compared with outcomes that would have resulted without the program but with all other conditions identical. In ESAA, by contrast, we have evidence that the existence of ESAA in the treatment schools affected conditions in the control schools. In particular, we are quite certain that many district administrators reallocated other special funds, such as Title I money, to the control schools to compensate for their not receiving ESAA funds. (This reallocation, it should be noted, was both legal and ethical despite its unfortunate effect on the evaluation effort.) Furthermore, it appears that some control schools used the extra funds to support activities similar to those supported by ESAA. Thus, although we had rigorous control over the assignment of schools to ESAA and non-ESAA conditions, our control over total school dollars and activities was considerably eroded. As a result, we have tended to place greater emphasis in our first-year ESAA report on outcome differences for different types of program characteristics, regardless of how these programs were funded, then on the ESAA/non-ESAA distinction.

In studies where random assignment to treatment and control conditions is totally impractical, other means must be sought to create appropriate comparison groups. In the study of Sustaining Effects of Compensatory Education, we hope to find students who are not receiving Title I aid, but who are otherwise comparable to the Title I students. We anticipate that this will be a difficult job, because of the large proportion of disadvantaged schools and students receiving Title I benefits. Certainly there is reason to expect that non-Title I schools, even those in the lower end of the socio-economic spectrum, may be systematically different from Title I schools on important variables other than Title 1 participation. The design problems inherent in such a situation have been convincingly pointed out by Lord (1967) and by Campbell and Erlebacher (1970), especially for the case in which pretest means differ from the treatment and control groups. A later portion of this paper discusses techniques by which we hope to derive legitimate outcome comparisons in the Sustaining Effects Study.

Still another approach in seeking comparison groups for program evaluation is the use of before-and-after data. For example, achievement gains before a program was introduced may be compared with those measured after program implementation. The validity of such comparisons is seriously threatened in most studies, however, by the many uncontrolled non-experimental differences that may exist between the two time intervals being contrasted.

A variant on the before-and-after comparison is the use of national norms. Based on test publisher's norms, for example, one can determine the mean pre-post gains associated with a national sample of students having the same distribution of pretest scores as those in the treatment group. One can then test whether treatment group gains were larger than those made by the hopefully comparable normative group.

In the ESAA study, using the publisher's data for the California Achievement Test, we found that sample students in both ESAA-funded and non-ESAA schools made larger gains than would have been expected on the basis of national norms for low-achieving students--that is, for students with entry scores similar to those of students in districts receiving ESAA funds. Our interpretation of this finding is that ESAA funds in a district tended to increase total funds and associated resources in both the ESAA and non-ESAA schools, thereby having a generalized positive impact. This interpretation

is tempered by a number of methodological constraints, one of which is the fact that the California Achievement Test, like most other standardized achievement tests, has empirically based normative data for only one point of time in each school year. Normative data for all other points in the school year were interpolated by the publisher on the basis of largely unsubstantiated assumptions about the shape of the growth curves over the summer and during the school year. Any serious inaccuracy in those assumed growth curves could obviously invalidate use of the normative data for benchmark purposes.

In the multi-year Sustaining Effects Study, we have an opportunity to develop our own norms for the California Test of Basic Skills, based on a nationally representative sample of over 150,000 students. This sample will include both Title I and non-Title I schools and students covering the full spectrum of socio-economic status and racial/ethnic composition. Also, tests will be given twice a year so that empirically based fall and spring norms can be developed for each grade level. While the shape of growth curves within a school year will remain somewhat conjectural, it will be possible to determine what happens to achievement over the summer months. Furthermore, we will obtain fairly complete demographic data on each student and school in the sample, so that characteristics of different comparison subgroups can be accurately specified. Thus, we expect to be able to place considerable confidence in the normative data and to have a defensible benchmark for assessing the achievement gains of Title I participating students.

Finally, the special case of criterion-referenced testing should be considered. Where absolute criteria of performance have been operationally defined, the need for comparison groups or norms may seem totally eliminated. Either the students in a program meet the performance standards or they do not. When viewed more closely, however, this hard-and-fast distinction seems to soften somewhat. To begin with, meaningful standards or performance criteria are not that easy to define in the abstract. How does one decide exactly what mathematics skills every student should have at the end of the third grade? How does one justify setting a particular requirement for that level rather than for the second or the fourth grade? Such absolute criteria are much more easily established for a specific

instructional package or one-shot training course than for complex, multi-year and multi-faceted programs such as Title I. In the more complex programs, if the various stages of objectives and criteria are not to be totally arbitrary, it is likely that they will be based, in part at least, on internalized normative standards. That is, one's prior experience with the accomplishments of students at a particular school level will probably be a major determiner of one's criteria for gauging future students' success.

In general, the notion that criterion-referenced testing eliminates the need for normative standards or comparison groups appears more valid for individual student assessment and for formative program evaluation than for summative evaluation of program impact. Perhaps the purest and most effective application of criterion-referenced testing is its use in a prescriptive role, that is, in determining a student's readiness for exposure to a more advanced level of instruction or his need for further review of an earlier concept. In a similar sense, during program development, criterion-referenced testing can effectively reveal specific areas of deficiency in the program's content or sequence. In summative program evaluation, by contrast, one is rarely content with knowing simply that one or more students have or have not successfully met some prespecified performance criteria. More often, the program being studied is explicitly or implicitly in competition with another program or mode of instruction, and the issue of relative effectiveness is important. Such an evaluation, therefore, involves, once again, a comparison of results under alternative treatment conditions. This comparison may be based on the percentages of students reaching a given criterion in a specified time period, on the amount of time required for a certain percentage of students to achieve a criterion level, or on some combination of such factors. The point is that most large-scale evaluations of social programs have at least a large summative component, and such evaluations typically require norms or comparison groups, even if they use criterion-referenced tests.

The Need for an Integrative Evaluation Strategy

The remainder of this paper steps back, in a sense, from particular issues of evaluation methodology, and considers more broadly the role and general approach of education evaluation, especially in relation to more traditional types of education research. A general strategy is described--or, more accurately, a tentative outline of a strategy is suggested--by which future evaluation studies might be guided so as to increase their contribution to our generalized knowledge of educational processes.

It has frequently been pointed out that there are fairly distinct differences between research and evaluation, and that the two terms should not be used interchangeably. An evaluation is designed to provide information about the effectiveness or impact of some specific procedure or program, when that procedure or program is applied in a designated context. Results of evaluations are usually intended to aid in decisions concerning the desirability of adopting or continuing the program in question. For this reason, there is generally greater interest in guaranteeing that the evaluation provides a fair and realistic test of the particular program, than in discovering general principles that might be applied to many programs in many different settings. In part as the result of this emphasis on realism, though certainly not as an intentional design feature, it is sometimes difficult to control or even to describe all relevant contextual variables to a degree that would allow complete replication of the study conditions.

Research, by contrast, seeks to build a systematic body of knowledge that, within defined limits, can be applied to a wide variety of circumstances. Its goal is not to aid in some immediate decision about a specific program or procedure, but to uncover and explicate patterns of associations among variables of interest. To achieve this goal, research stresses replicability, and attempts to manipulate the experimental variable(s) of interest while holding other contextual variables constant.

In principle, the above statements contrasting research and evaluation are largely true. Certainly the popular conception is that evaluation may provide practical answers to immediate administrative information needs, but that research is the path toward eventual understanding of complex interactions

in the teaching-learning process. In practice, however, the presumed differences between research and evaluation do not always seem that clear cut. For one thing, applied research often falls somewhere between basic research and evaluation. To cloud the distinction still further, the term "evaluation research" has been recently applied to studies that incorporate in an evaluation context many of the goals and methodologies of research. Whatever labels are used, the important point is that evaluations appear to have the potential for contributing at least as much to our basic understanding of education and of the teaching-learning process as classical research. (Of course, many skeptics do not regard classical research's achievements in education as having set very high standards for competition. At the same time, however, it seems premature to accept Cronbach's (1975) pessimistic view that it will never be possible to achieve "enduring systematic theories about man in society.")

The remainder of this paper presents arguments for the thesis that evaluations have potential for significantly increasing our generalized knowledge of teaching-learning interactions, and describes one possible strategy for realizing that potential.

First, several of the major evaluations over the past ten years or so have been conducted on a scale, and with a sophistication of design exceeding that of most traditional, research-oriented studies in education or psychology. Some of these evaluations have involved not just one or two classes, or even a few schools, but literally scores of districts, hundreds of schools, and many thousands of students. It seems fairly safe to conjecture that three or four of the recent large-scale evaluations have collectively included more students than the total of all basic research studies reported in educational and psychological journals over the past ten years. Furthermore, this trend is likely to continue, at least for the foreseeable future, since evaluation requirements are being built into the legislation for a number of the new Federally aided programs.

Of course, bigger is not necessarily better. A superficial survey or poorly controlled study, however large, may advance our understanding of behavior much less than a well designed but small-scale experiment. But evaluation

methodology itself has matured considerably over recent years, and some of the latest evaluation designs have begun to compare favorably with the more classical research paradigms in attempting to control or at least reduce sources of contamination and bias in the findings. True, it is rarely feasible in evaluations to hold virtually all variables constant, as in traditional experimentation, but advantage can be taken of the naturally occurring variations, by recording those variations and examining interactions among the variables. It is this realm of interactional or relational analysis, in fact, where current evaluation methodology may have some distinct advantages over basic research, at least as basic research has been conducted in the past with its emphasis on simplification and control. In our ESAA evaluation, for example, we explored possible linkages among district-level, school-level, and teacher-level variables intended to influence student achievement. One of the more interesting findings was a positive relationship between the schools' desegregation-related activities (e.g., parent-staff conferences) and students' achievement scores; in addition, more positive student attitudes toward school and toward their own educational futures were found in schools that had greater amounts of student interracial activity, more positive teacher and superintendent attitudes on desegregation issues, and more frequent desegregation-related district and school activities.

Partly because of the interest in reporting descriptive data, and partly because of the desire to study interactions among student, context, program, and outcome variables, large-scale evaluations tend to collect data on a much broader and richer range of variables than is typically true of basic research studies. These data often include, for example, information about the community and district environment, staff and student backgrounds, physical and budgetary resource allocations, administrative policies, and overall program operations, as well as more specific data on classroom interactions and student achievement. Furthermore, some of the larger evaluations include repeated data collections over a period of several years, thus permitting longitudinal analyses of cumulative impact, fade-out effects, age-related effects, etc. Traditional experiments, by contrast, seldom span more than a single school year, and often involve much briefer periods.

This is not to say that evaluations can or should replace basic research, but simply that they share with research the potential for increasing our generalized knowledge of educational processes. So far, unfortunately, this potential has not been realized to any great extent. There are few general principles of learning-teaching interaction that one can confidently state on the basis of evaluation studies. The ad hoc and ideosyncratic nature of evaluations has made it extremely difficult to synthesize results across those evaluations in any meaningful way. Each evaluation has dealt with somewhat different program and context variables, with students differing in age level or other characteristics, and with different outcome measures. To some extent, of course, these differences are inevitable; evaluations are, by definition, tailored to the specifics of the program being evaluated. Nevertheless, it should be possible to develop an overall evaluation strategy using a conceptual framework or model of educational process to help guide the design of new evaluations. Results of the new evaluations could, in turn, be used to refine and extend the conceptual framework.

The use here of the term "conceptual framework" rather than "theory" is deliberate. While many investigators find it useful to invoke theoretical constructs, it is by no means obvious that such constructs are an essential ingredient in developing or applying effective evaluation strategies. What does seem essential is some form of roadmap or matrix by which it is fairly easy to determine the status of evaluation in a given area of application. This matrix should show in a systematic, organized fashion what evaluation policy issues have been addressed in previous stidies (or are currently being addressed), what sets of variables have been studied in relation to each other, what range of values has been used for each variable, and what patterns of association have been empirically identified. Then, as new evaluations are being planned by a funding agency such as USOE, each could be compared with the matrix of past and current studies to determine how the new study might best complement and extend the existing knowledge. Efforts would then be made to set design parameters for the new evaluation so that it will not only satisfy the immediate evaluation requirements, but will also contribute maximally to the conceptual framework which represents science's understanding of educational processes. In some cases this contribution might involve

replication and verification of findings made in other evaluations; in other cases the new study might help to fill gaps in the variables and relationships that have previously been examined.

The evaluation strategy discussed here represents an effort to improve the design and integration of new studies, and to present interrelationships among evaluation findings more clearly than is possible today. It does not represent an attempt to dictate what programs should be evaluated, although the existence of an effective and frequently updated conceptual framework would presumably be helpful in deciding on evaluation priorities. In this sense, the present suggestions are considerably less comprehensive and ambitious than those of Bernstein and Freeman (1975), for example, who have recently recommended quite drastic changes in the entire administrative mechanism for selecting, awarding, and monitoring evaluation studies across major Federal agencies. At the same time, the integrative evaluation strategy discussed in the present paper may have more immediate applicability, since it can be applied without major changes in the national power structure. Congress has made it quite evident that it intends to exercise greater discretionary control, not less, in specifying its own mandates and priorities for program evaluations. Furthermore, within those Congressional mandates and priorities, administrative agencies such as USOE and NIE have shown increasing desire and competence to make their own judgments in selecting the programs to be evaluated, and in setting the basic evaluation designs. Thus, there seems little reason to anticipate revolutionary changes in the near future in the basic mechanisms for structuring the evaluation requirements; instead, this paper is concerned with strategies for optimizing the contributions of evaluations, and with methods of applying those strategies within the existing administrative mechanisms. Even when certain basic design features of an evaluation have been fixed, there are usually many other, more detailed design decisions that could be made in one direction about as easily and defensibly as in another direction. These decisions may involve, for example, what types of background and contextual variables should be examined, and what specific instruments should be selected to measure program outcomes. It is at this discretionary level that an evaluation strategy might be applied almost immediately to enhance the overall utility of new evaluation studies and to integrate their findings more effectively.

There is ample evidence, it should be noted, that some federal agencies, at least, are fully aware of the need to build on an existing structure of findings from previous studies, rather than to treat each new evaluation as in a vacuum. In several recent evaluation procurements from USOE's Office of Planning, Budgeting and Evaluation, for example, there are visible efforts to verify preliminary results of earlier studies, to extend the range of grade levels examined in the earlier evaluations, and so on. What appears needed is a more systematic, perhaps a more formally structured, and certainly a more widely communicated attempt to apply an integrative evaluation strategy in all agencies dealing with education-related evaluations.

This brings us to the question of who should develop and implement the conceptual framework and evaluation strategy discussed here. There seems little doubt that, under the present administrative structure, at least, implementation must be in the hands of the Federal agencies designing evaluations, reviewing proposals, and awarding evaluation contracts and grants. For the general field of education, this means primarily USOE's Office of Planning, Budgeting, and Evaluation (OPBE), the National Institute of Education, and the National Center for Educational Statistics, though other agencies do sponsor evaluations from time to time. At a higher level, of course, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) in HEW oversees the total evaluation picture across all of those agencies. But, technical decisions of the type most immediately concerned here, such as selection of specific program and context variables, are usually made within OPBE and similar agencies; therefore, this level--the level of line organizations, so to speak--may be the best starting point for practical application of an integrative evaluation strategy. At the same time, there would obviously be major benefits if those line organizations interacted closely in refining the conceptual framework and updating it at frequent intervals to reflect new evaluation and research results. It is quite possible that ASPE could serve an effective role in coordinating those interactions.

In principle, those same Federal agencies could also develop the conceptual framework and evaluation strategy discussed here. Unlike the situation five years ago, there now appears to be more than adequate talent and

expertise within the ranks of HEW to prepare excellent models for guiding an integrative evaluation design. The real question is whether the line organizations could afford to release the key personnel from their ongoing administrative and monitoring responsibilities for long enough to do the required conceptualization and development. If adequate resources cannot be spared, it might be valuable for the government to appoint an advisory panel consisting of national leaders in educational evaluation design and methodology. The panel's role would be to work with federal staff members in preparing the conceptual framework. To be effective, such a panel would need to be fairly small and highly product-oriented. An excessively cumbersome and slow-moving group, no matter how prestigious its members, would simply interfere with the ongoing schedules and administrative chores of the federal agencies, and probably retard rather than advance the cause of educational evaluation.

REFERENCES

Bernstein, I. N., and Freeman, H. E., Academic and Entrepreneurial Research, New York: Russell Sage Foundation, 1975.

Campbell, D. T., and Erlebacher, A., How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory evaluation look harmful. In Jerome Hellmuth (Ed.), Compensatory Education: A National Debate, New York: Bruner/Magel Inc., 1970.

Cronbach, L. J., Beyond the two disciplines of scientific psychology. American Psychologist, February 1975, 116-127.

Lord, F. M., A paradox in the interpretation of group comparisons. Psychological Bulletin, 1967, pp. 68, 304-305.